Available online at www.sciencedirect.com

ScienceDirect

C. R. Physique 7 (2006) 896–909

COMPTES RENDUS

PHYSIQUE

http://france.elsevier.com/direct/COMREN/

Ultimate lithography/Lithographie ultime

Advanced mask manufacturing

Carlo Reita

CEA/LETI, Minatec, 17, rue des Martyrs, 38054 Grenoble cedex 09, France

Available online 21 November 2006

Abstract

In this article the fabrication of advanced masks will be briefly outlined and some specific aspects will be discussed. It will be shown that the increasing design complexity and the complexification of the optical lithographic process generate major issues on all aspects of the fabrication and control of such masks. These issues have not only technical but also economical impacts that will also be outlined. ***To cite this article: C. Reita, C. R. Physique 7 (2006).***

Résumé

Fabrication de masques avancés. Dans cet article nous esquisserons rapidement la fabrication de masques avancés et certains aspects spécifiques seront discutés. Nous montrerons que la complexité croissante des circuits et la complexification des procédés de lithographie optique génèrent des problèmes majeurs sur tous les aspects de la fabrication et du contrôle de tels masques. Ces problèmes ont non seulement des impacts techniques mais aussi économiques qui seront aussi indiqués. ***Pour citer cet article : C. Reita, C. R. Physique 7 (2006).***

Keywords: Mask; Mask writing; Phase shift mask; Mask inspection; Mask repair

Mots-clés : Masque ; Écriture de masque ; Masque à décalage de phase ; Inspection de masque ; Réparation de masque

1. Introduction

Photolithography masks, also called for historical reasons ‘reticles’, are a fundamental part of the production of integrated circuits although an often hidden one. They are the template by which the design of a layer of an integrated circuit is transferred onto the wafer. The relative facility of producing such masks until the 0.25 µm technology node has generated in the microelectronic industry the idea that mask is an easily manufacturable consumable. Unfortunately, the introduction of sub-wavelength lithography put suddenly a very high burden on masks, whose commercial manufacturers, especially in USA and Europe, were not ready to tackle. This concern originated from the mutual interaction of various factors, namely new type of masks, finer structures to be patterned especially to compensate for the optical degradation of the image in the sub wavelength regime, a large increase of the amount of data to handle due to these additional features and to the increasing density of the designs.

In this article we will discuss the entire chain of fabrication of advanced masks and the associated technical and economical issues. The concepts of advanced lithography, like Reticle Enhancement Techniques (RET), Phase Shift

E-mail address: carlo.reita@cea.fr (C. Reita).

1631-0705/$ – see front matter © 2006 Académie des sciences. Published by Elsevier Masson SAS. All rights reserved.
doi:10.1016/j.crhy.2006.10.006

Masks (PSM), etc. will be discussed only in relation to the mask production, a more detailed explanation of their use and their characteristics can be found in the other papers of this issue. We will also assume that the reader is familiar with today's photolithographic process. A more in-depth treatment of the subject of this paper can be found a recently published book [1] which is an extremely useful reference. The outline will closely follow the real production flow of a reticle: we will start with the general description of a mask, followed by the analysis of the treatment of the data necessary for the fabrication, then the fabrication sequence—including resist patterning, transfer of the image into the appropriate layers by etching, cleaning—, the metrology of the features, defect inspection and repair, and finally the pelliculation. A specific section on EUV masks will then describe some of their specific aspects and finally an analysis of the economics issues of the mask industry will be given, together with a possible view for the future.

2. The mask

The microelectronic industry has standardised all the parameters of a mask [2]: the substrate is a square of side 6 inches, with a thickness of 0.250 inches, on which a pattern is defined by etching in an absorbing and/or phase shifting layer deposited on its surface. The pattern is protected by a polymeric pellicle mounted on a frame attached to the substrate (see Fig. 1). For the critical layers of sub-100 nm technologies, the reduction factor is ×4 associated with a 193 nm lithography, but it is important to note that the largest volume of masks is still produced for 248 nm and 365 nm systems some of which use a ×5 reduction. The reason is double: (i) there are still a large number of new designs produced in older technologies; and (ii) at the same time for nodes below 120 nm only 10 to 15 layers out of the 30 plus layers required are really critical and require a shorter wavelength.

The material used for the substrate is most commonly fused silica, but can be some borosilicate glass for non critical layers. Historically the substrate has always been seen only as a mechanical component. Its only optical specification was the transmission at the wavelength used. More recently, other parameters showed some impact on imaging, in particular the residual stress birefringence, as demonstrated for the 157 nm lithography [3]. The current introduction of polarisation control in dry and immersion lithography will likely require this parameter to be specified. The planarity of the surface also becomes an increasingly important parameter for imaging: its value is presently comprised between 0.5 and 2 µm.

The present fabrication of a mask starts from a so-called 'blank' i.e. a substrate coated with the appropriate material stack and the top resist layer. In Fig. 2 are shown the schematic cross-sections of different masks blanks types. Standard binary masks (Fig. 2(a)) use Cr as absorber mostly for historic reasons, although there are indications that an alternative material would be preferable both from an optical and a processing point of view. Normally the absorber is covered with an antireflective (AR) layer to reduce flare in the imaging system. This AR coating is usually an abrupt or graded layer of chromium oxide or oxynitride. The thickness of the absorber layer, as well as the details of the AR coating, vary depending on the wavelength for which the mask is fabricated, either 248 nm and above or 193 nm. The major specification is that the absorbing stack must have an optical density equal or higher than 3: so the currently available thicknesses are about 100 nm for 248 nm and above lithography and 70 nm for 193 nm. More recently an optical density of 2.5 has been deemed acceptable and thicknesses of 50 nm are available. The thickness has an impact on the optical performance because the assumption that was usually made of an ideal optical 2D opening is increasing

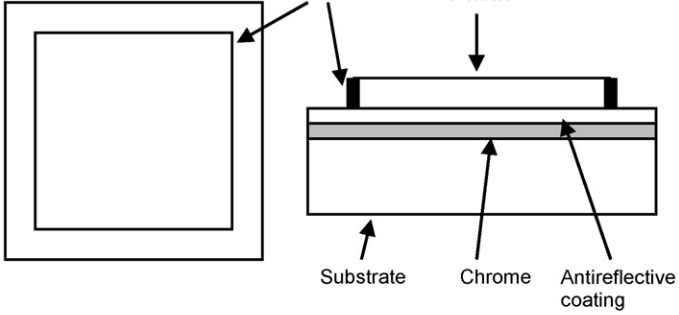

Fig. 1. Schematic view of a mask blank and frame mounted pellicle (sizes not to scale).

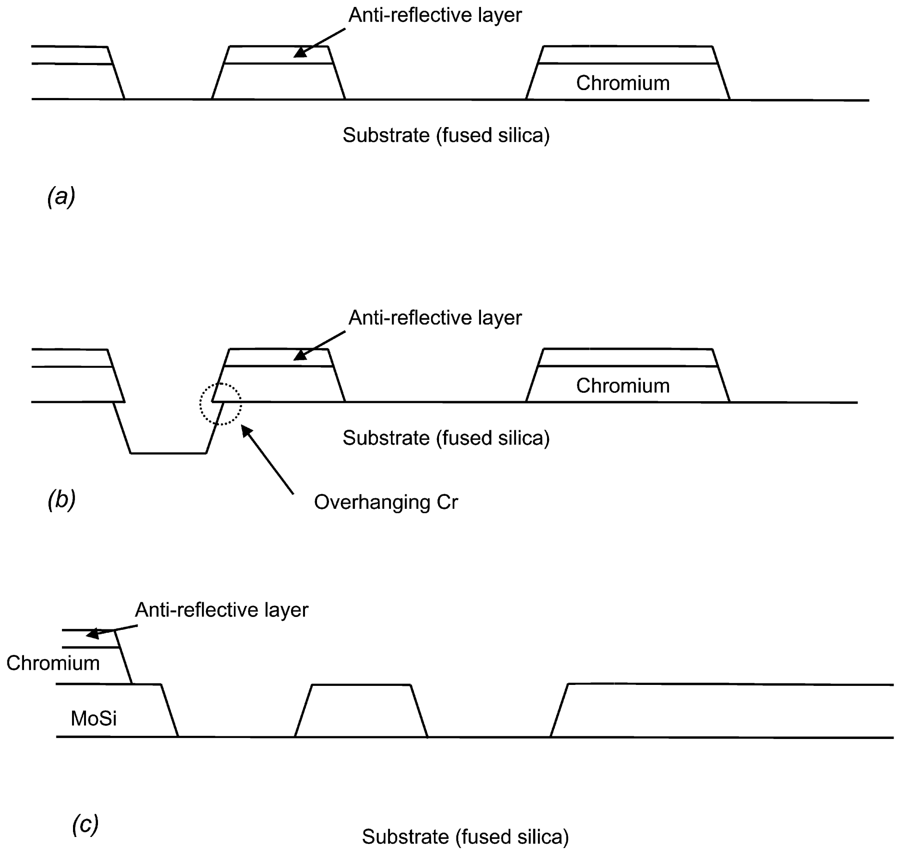

Fig. 2. Schematic cross-sections (a) of a binary mask, (b) of an Alternating Aperture Phase Shifting Mask (AAPSM or AltPSM) and (c) of an Embedded Attenuated Phase Shifting Mask (EAPSM or AttPSM).

questionable as the thickness and the features on the masks start to be comparable in size. In turn this implies that optically the mask is no longer a 2D object but a real 3D one.

The Alternating Aperture Phase Shifting Mask (AAPSM or AltPSM) shown in Fig. 2(b) is also obtained from a conventional binary blank by etching into the fused silica substrate using Cr or the resist as masking layers. On the other hand, the Embedded Attenuated Phase Shifting Mask (EAPSM or AttPSM) shown in Fig. 2(c) requires a phase shifting layer usually inserted between the substrate and the absorbing layer. The most commonly used material for such a layer is MoSi and, depending on the final wavelength of use, the thickness and stoichiometry of this layer determine the transmission characteristics. The most common transmission values are 6 or 9%; 18% or more has been investigated but is less common and not commercially available. For AttPSM masks, the presence of the phase shifting layer reduces the absorption requirements of the absorption layer, which is then thinner.

3. The data preparation

The pattern to be defined onto the mask is today drawn using modern CAD systems by large teams of designers who release the drawings in an electronic form. For over thirty years now the interchange format for the drawings has been Graphic Design System II (GDS2), originally created as an internal format by the Calma Company, but then so widely accepted to become a de-facto industry standard: its simplicity and versatility has made its success. This format represents the polygons that form the layout in a binary vector data system. Its two main characteristics are the possibility to represent arrays of features and the possibility of re-using defined structures in a hierarchical form. This has been sufficient for a long time but, today, the introduction of RET and the design complexity have made necessary the definition of a new format. This has been done by a SEMI committee that has created the Open Artwork System Interchange Standard (OASIS) format which is progressively been introduced. This new standard takes advantage of the new computing capabilities (e.g. 64 bit addressing) and compression techniques to maintain manageable data sizes [4].

In order to fabricate the mask, however, the drawing of the circuit layout as produced by the designers is not sufficient. It is then necessary to convert the different patterns into a language specific to the mask writing tool. The different patterns are then assembled as needed on the surface of the mask. Finally the inspection and the metrology files for the controls during and after manufacturing are generated and added. The ensemble of these operations is normally referred to as Mask Data Processing (MDP).

Most of the different patterns which are on the final mask are defined by the user and can be one or multiple instances of the same die, multiple layers (for what are called Multi Layer Reticles, MLR), control patterns, scanner specific alignment patterns, barcode for automatic handling etc. In turn, these patterns may require extra processing, the more common case being that of a layer not drawn but derived via Boolean operations from other layers (e.g. masks for implantation layer). They all need to be translated into the format specific to the mask writing tool, essentially reduced to trapezoid on the addressing grid of the machine. This process is normally referred to as 'data fracturing'. This step is one of those that are becoming increasingly difficult at each new technology node: the combination of the design size and of the introduction of optical correction features implies a huge increase in the number of elementary trapezoids in which the design is decomposed and hence in the size of the associated file. For a 90 nm design, for one layer, this size may already exceed 40 GB. It tends to increase more than quadratically at each new design node, although efforts are made to contain this increase by better designs and practices.

Once all the patterns are generated, they have to be assembled. Besides the already indicated features it is necessary to add the patterns needed by the mask manufacturer as alignment marks for different writing tools, control patterns, titles, logos, etc. This phase is called 'Job Deck' because it takes the shape of a series of instruction (a 'deck') on how to place and size all the patterns.

It is worth to point out that these two phases are probably the most critical in manufacturing, as any error in this data generation and handling may cause the mask to be incorrect and the error could, in the worst case, be spotted only after full wafer fabrication at electrical testing level with huge economic consequences. At the same time these are the phases where human error is more likely. Also, issues may arise not only as a consequence of a mistake, but just by the transformation of the design from the design grid to the grid specific to the writing tool that are not necessary multiples. An example for what can happen is shown in Fig. 3 and to reduce to a minimum the occurrences of such deformations a large number of controls need to be put in place at this stage.

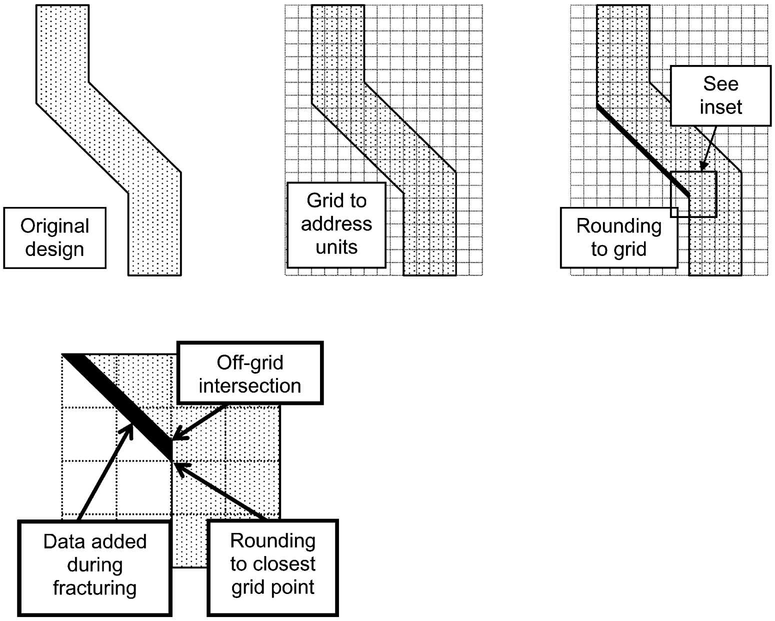

Fig. 3. The transformation of the design from the design grid to the grid specific to the writing tool in the data fracturing may induce possible feature distortions induced by the rounding process.

The last step is the generation of all the files necessary to perform the metrology and inspection on the fabricated mask according to the specific customer requirements. This task is again time consuming and would benefit from an increase in automation that is hampered by the lack of data exchange standards between tools. After this stage all the data are ready to start the manufacturing process.

4. The pattern definition

The actual definition of the pattern onto the mask is done via direct writing onto a resist either by using a laser or an electron beam (also called e-beam). Each of the two methods has its advantages, but the resolution requirements of advanced masks is such that today all masks with features smaller than about 300 nm at mask level (sometimes referred as 'at 4x' to distinguish from the size on wafer called 'at 1x') are made by e-beam. However, for completeness, the main characteristics of both systems will be described, as the work on new laser writers could bring them to be again competitive for critical layers.

Advanced laser mask writers are all based around multi Gaussian beams that are raster scanned on a finite grid across the surface of the blanks. The beams are modulated electro-optically and can be controlled in up to 64 grey levels on the most advanced tools. Today two families of equipments exists, those using a linear array of beams (up to 32 beams on the ALTA 4700 from Applied Materials) and those using a rectangular array (up to 512 × 2048 on the Omega 7300 from Micronic), the latter having a larger throughput but needing a more complex data handling architecture. These systems use mainly DUV lasers (e.g. 248 nm for the Omega or 257 nm for the ALTA) compatible with optical standard and Chemically Amplified Resists (CAR) already widely used in wafer manufacturing. The maximum resolution is determined by a combination of the wavelength, the numerical aperture of the system, the number of grey levels and the writing strategy. Limiting ourselves to the two already mentioned tools, the announced lithography performances are reported in Table 1. As already said, the main advantage of lasers is the high throughput coming from the use of multiple beams that reduces writing time. The second factor is the relatively lower complexity when compared to an e-beam writer (no vacuum, lower sensitivity to EMI, no charging issue associated with the substrate, etc.) which makes the cost of ownership (CoO) of such tools considerably lower. For this reason they tend to be used as long as their resolution is acceptable and a large interest exists in further developments.

E-beams, however, remain the only tools today capable of producing the resolution required in the critical layers for current and possibly future technology nodes. A factor that is still not widely recognised, in fact, is that the requirements of RET features make the resolution requirement on the mask similar to those on the wafer. Today, at 65 nm node, it is quite common to have 100 nm wide features on the masks (e.g. scatter bars or scatter trenches) as well as 40 nm edges (e.g. jogs or hammerheads) for which a high pattern fidelity is required. In Fig. 4 it is shown how the application of an Optical Proximity Correction (OPC) affects the shape of a feature on the mask and makes it more complex to manufacture and control. At the same time the incremental improvement in pattern fidelity on the wafer by applying increasingly complex OPCs is progressively diminishing inducing some necessary trade-off.

Current e-beam tools are all operating at 50 keV beam energy. This parameter will probably not change in the future due to the cost of developing new equipment and due to the reduced resist sensitivity at the higher energies that would improve the resolution. The e-beam tools can be divided in different classes according to the scanning techniques (raster scan or vector scan) and the beam shape (Gaussian beam and shaped beam). The paramount requirement of throughput has made the vector scan, shaped beam the tool mostly used today.

Table 1
Characteristics of two commercial laser mask writing tools

Specifications	ALTA 4700 (Applied Materials) (nm)	Sigma 7500 (Micronic) (nm)
Minimum main feature	260	220
Minimum assist feature	140	130
Address grid	1.25	1.25
CD Uniformity (global 3σ)	6	5.5
Registration (global 3σ)	8	12
Second level alignment	20	20
Wavelength	257	248
Resist process of record		FEP 171

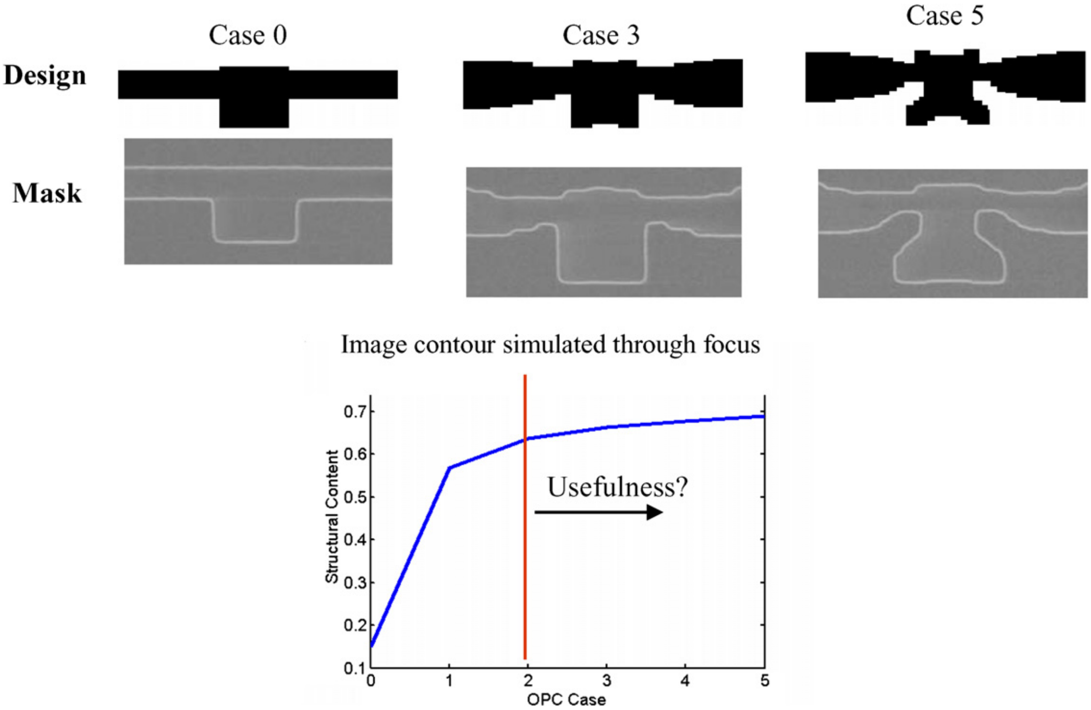

Fig. 4. Application of progressively more complex Optical Proximity Corrections (OPCs) to a feature. The mask becomes increasingly more complex to manufacture (more rectangles, smaller sizes) while the results on the wafer becomes relatively less significant beyond a certain OPC level. (Courtesy of C. Progler—Photronics Inc.)

The process of e-beam writing of masks is more complicated than that of direct writing on wafer, mainly as a consequence of the thermally and electrically insulating substrate. The thermal isolation makes the pre- and post-bake of the resist a very critical process, especially for e-beam CAR resists whose dimensional sensitivity is around 0.1 nm per degree Celsius. Furthermore the resist heating during e-beam exposure has an impact on the imaging performance. The electrical insulation induced by the fused silica blank and the AR coating leads to a progressive electrical charging of some areas: variations in the deposited dose and in the local beam positioning impacts the final critical dimensions (CD), their uniformity and the local positioning of patterns (usually called ‘registration’). These charging effects are exacerbated when writing the second or third levels of a PSM in which the Cr layer is no more continuous. The combination of heating and charging determines an upper limit to the current density of the electron beam and hence a lower limit to the writing time.

Another important issue with e-beam writing is the electron scattering in the resist and in the substrate, in particular the back-scattering. The back-scattering has a global component depending on the stack composition, and a local one depending on the type and density of the features. In fact the backscattered electron constitutes a dose background that can locally alter the CD as shown schematically in Fig. 5. To compensate for these effects all production tools provide, in their data treatment system, provision to apply a Proximity Error Correction (PEC). The algorithms and methods vary among the vendors but the principle is the same: the patterns are ‘binned’ to determine the local average amount of backscattering and then to calculate the dose correction to be applied to each bin. This process is done ‘on-the-fly’, i.e. during the writing, in order to speed up the overall writing time; however, it adds to the complexity of the electronics handling of the data. It is worth noting that on some tools the PEC is done using analogue circuit boards to allow as much real time computation as possible and to reduce overheads.

After irradiation the resist goes through the post-exposure bake and development. As already mentioned, this step is more critical than the corresponding one in wafer processing due to the thermal insulating nature of the substrate, the intrinsically longer delay time from the end of the exposure to the start of the process (return from vacuum to atmosphere, manual transfer, etc.) and the uniformity issues (square substrate). The use of CAR resists, necessary to maintain an acceptable throughput, has also a negative side effect due to the increased line edge roughness related to

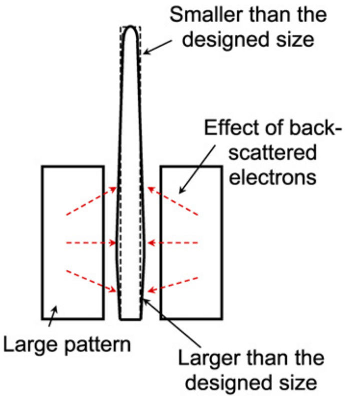
Smaller than the designed size
Effect of back-scattered electrons
Large pattern
Larger than the designed size

Fig. 5. The backscattered electrons in an e-beam mask writer induce a variation in the actual dose locally received by the resist. This dose variation is dependent on the local pattern arrangement and has to be corrected by Proximity Error Correction algorithms.

the acid diffusion. Today this roughness is of the order of 5 nm per edge and, unless brought under better control, will soon constitute a limit to the CD tolerances.

Once developed, the pattern can be etched first into the Cr. The Cr etch is performed by dry plasma processing using Reactive Ion Etching in chlorine based mixtures. Historically Cr was chosen for its very good durability and the ease of wet etching it. When the need for a higher resolution imposed to switch to dry etch, problems started to appear as Cr forms very few volatile compounds. Other materials have often been proposed but never acted upon by commercial blank suppliers. The major issue with the dry etch is the low selectivity between Cr and resist in all Cr dry etch chemistries that induces a variation of the resist feature size during the process. This is partially compensated during the data preparation by appropriate sizing of the patterns for a given process, but this technique becomes increasingly difficult with the reduction of feature size as corrections are approaching 50%. Another issue is the local variation of the etch rate depending on the pattern density, this effect being due to the different concentrations of species related to the higher or lower real percentage of material exposed. An optimum of the final CD uniformity is reached by combining a sizing of the features at the data preparation step, an optimisation of the PEC parameters and an adjustment of the dry etch parameters.

In the case of EAPSM and AAPSM a second writing step is also necessary and laser is preferred to avoid charging issues. However some very advanced masks, like those required in zebra Chromeless Phase Lithography (CPL) [5], due to resolution constraints, require a second e-beam writing. After writing, the MoSi or the fused silica are etched to obtain phase shifting transmitting areas. In the case of MoSi the problem are similar to the Cr etch, although using a different chemistry, while the fused silica etch introduce a further problem namely the control of etching depth. In fact the etching depth in silica determines the phase shifting angle and needs to be tightly controlled: it translates into the fact that the density effects have an impact not only on the lateral dimensions but also on the phase.

As already mentioned, the reduction in features sizes makes 3D effects more pronounced and this is especially true for PSM masks. In Fig. 6 the wafer CD variation is shown for varying pitch and varying MoSi vertical edge slope of dense features [6]. It can be seen how, even with a correct feature size on the mask the CD variation on the wafer can be over 10% for slopes lower than 80 degrees. The 3D effects are even more pronounced in AAPSM masks due to the scattering from the sides of the trench etched in the fused silica. To obtain the necessary quality of imaging characteristics, part of the Cr should overhang the fused silica trench sidewall as schematically shown in Fig. 2(b). The process to control this Cr undercut is critical and is becoming increasingly so because feature sizes are approaching limits determined by the two conflicting requirements: the maximum overhang size is related to the mechanical stability of the feature, namely it should not to fall down during the chemical-mechanical cleaning and the minimum size is dictated by the effectiveness of the undercut (Fig. 7).

Cleaning is the last process of the actual fabrication and goes through multiple steps to remove organics and particles using different chemistries (acid or basic) and mechanical systems (which are mostly megasonics, but supercritical fluids, CO_2 snow and aerosols have also been tried). It should be noted that cleaning can introduce variations of phase and transmission in EAPSM masks due to the removal of MoSi: as a consequence there is a maximum number of times the cleaning process can be repeated during the lifetime of the mask before it fails out of specifications.

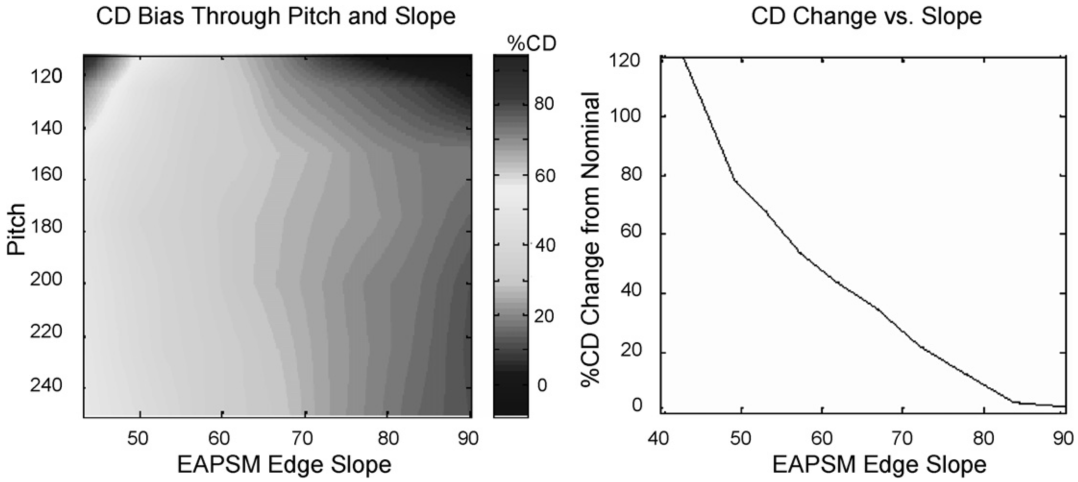

Fig. 6. Effect of the pitch and the mask edge slope on the wafer CD of dense features as obtained by optical simulations. (Courtesy of C. Progler—Photronics Inc.)

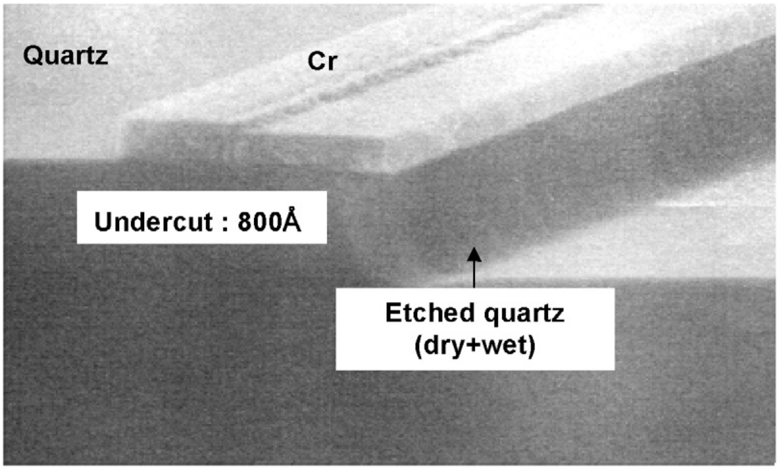

Fig. 7. Photograph of the cross section of an AAPSM mask showing the Cr overhanging and the fused silica trench (usually called quartz). The quartz etch has removed part of the anti-reflective coating (AR Cr). (Courtesy of Photronics Inc.)

5. The metrology

Once the reticle is finally fully fabricated, it is time to verify its conformance with the required dimensional, transmission and phase specifications. The dimensional one are the size and tolerances of critical features (normally called CD for Critical Dimensions) and the average positioning of the features compared to an ideal grid (normally called 'registration'). Fig. 8(a) illustrates a distribution of results corresponding to the measurement of a defined pattern on different locations on the mask. 'Target' is the expected size of the feature and the user defines the acceptable tolerance around this value. The actual results give the 'mean' value and the associated 'range'. The specifications are usually given in terms of a 'mean to target' and a 'range' or in term of 'maximum deviation' of any measure from the target. Fig. 8(b) stresses that tight distributions do not imply that values are within specification. As local variations in pattern density can induce process (write and etch) local variation, a nominally equal size feature can end up having a different distribution depending on it being close ('dense') or far ('isolated') from other features. The user can then specify the maximum difference between the two distributions. It can thus be possible that two very different distributions (badly compensated process) are within specification while of two close ones (well compensated process) are out of spec. It should be noted that the 'mean to target' difference in a stable process can be corrected by properly

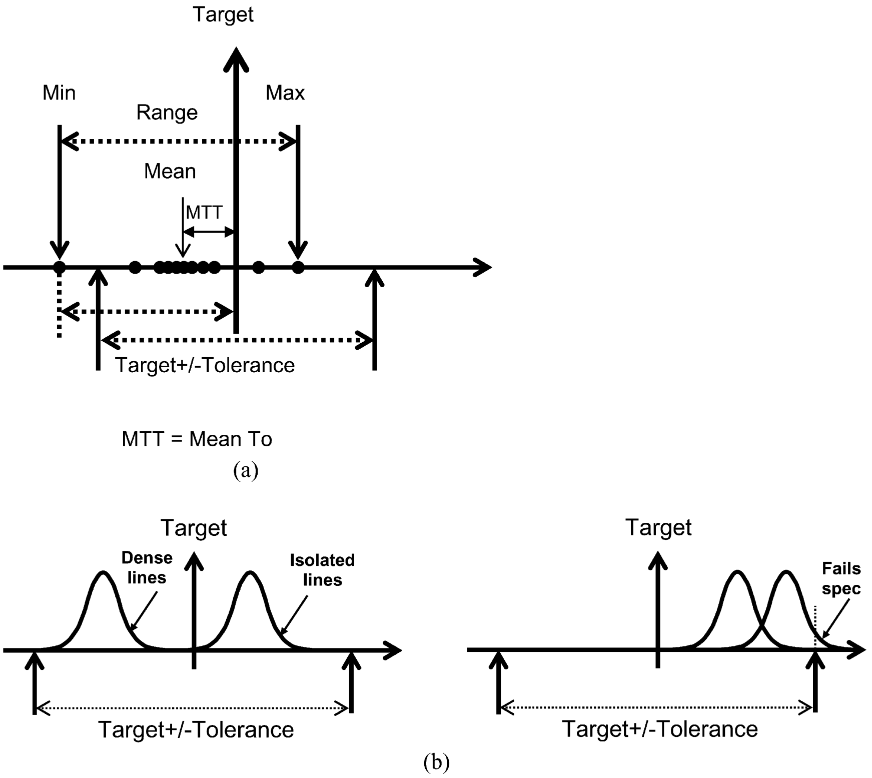

Fig. 8. (a) Definition of the terms used in CD definition and specification of nominally identical features measured across a mask. (b) Two different distributions of isolated and dense lines showing how more uniform distributions may still be out of specification while less uniform are not.

biasing the feature sizes during the data preparation, while the range is more difficult to control and is related to the uniformity of the process.

Traditionally the CD metrology was performed on specific marks present outside of the active die area and the number of sites to measure was in the few tens. Today, the criticality of the CD has increased due to the RET features and so it is necessary to measure a few hundreds of locations directly inside the active area with a consequent increase of the required time and the need of very good positioning accuracy and of pattern recognition capability. A very important issue is the lack of a recognised dimensional standard for the sizes currently of interest, i.e. down to 10 nm (the smallest currently available one stops at about 300 nm). For this reason the dialog between a mask supplier and a customer is based on a bilaterally agreed 'golden' standard and tables of correlation among tools at both ends have to be established through a time consuming and costly procedure. The resolution and precision required by the measuring system has also increased and typically today's systems have a requirement in a 3 sigma repeatability of about 2 nm and have to be able to resolve sub-50 nm structures.

The equipment used for routine testing is either deep UV microscopes or scanning electron microscopes (CD-SEM). Other techniques like Atomic Force Microscopes (AFM) or similar stylus type systems are used but only for process qualification and tool calibration due to their limited throughput. The optical microscope operates at 248 nm wavelength and its resolution limit is around 100 nm. The advantages of the optical tools are the similarity of the measurement procedure with the exposure mechanism in the scanner as both are performed with transmitted light, and its throughput and relative ease of setup (no vacuum). The CD-SEM has a higher resolution ($\sim$10 nm), but suffers from the problem of substrate charging that requires low ($<$600 V) acceleration voltages, hence long integration times. Furthermore the CD-SEM tends to induce a slight local variation of the mask transmission due to a surface modification induced by the electrons.

The major emerging issue today is the increasingly pronounced edge roughness of the features and the impact it has on the measurement results and on the correlation between what is measured on the mask and the result on wafer

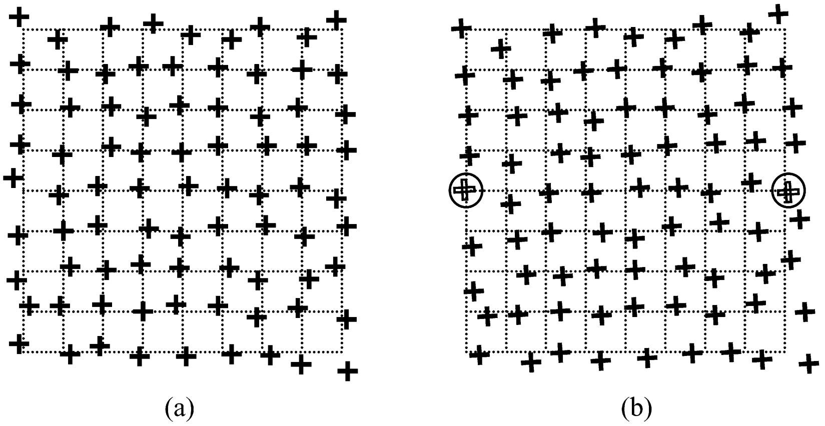

Fig. 9. (a) A Multi Point Matching registration is defined as the least square fit of all points so that the sum of the errors in both axes equals zero. (b) A Two Point Matching is defined in a similar way as multi-point matching except that the array is displaced and rotated to align two points with the 'design' grid before performing the least square fit.

after exposure. The algorithms employed and the measurement conditions have to be defined very carefully in order to perform meaningful analysis.

Apart from CD, it is also necessary to measure the distortion and the pattern placement compared to the ideal design grid in order to insure an acceptable superposition between masks of different layers. The registration is assessed by measuring on a specific tool the position of an array of alignment crosses placed at pre-defined locations outside the active area. In Fig. 9(a) and (b) are shown two different registration definitions that are currently used. The result is typically related to the performance of the writing tool employed and is slightly pattern dependent. As all substrate mountings during writing, measurement and utilisation are kinematics three-points suspension type, the measurement is a meaningful indication of the final performance. This is not true for EUV masks that are electrostatically clamped in the scanner and will very likely generate issues with the installed tool base when such masks will enter in full industrial use.

The phase and transmission characteristics are measured by optical interferometry, but given the size of the beams used in the tools the measurements have to be performed on specifically designed targets that are very large when compared to the real features whose characteristics may be slightly different.

6. Inspection, repair and pelliculation

Once the mask is certified correct in term of metrology, it is checked for defects. Defects are usually classified as soft or hard. Soft defects are particles of various natures on the surface or chemical contamination are usually larger than 1 μm and that can be removed by cleaning. Hard defects are printed defects in the pattern. These can be extrusion or intrusions of material on the edge of a pattern, pinholes or pindots in an opaque or clear area, or partially transmissive zones (thin chrome or thin phase shifting layer). It has to be noted that only those instances that would print on the wafer are considered defects, any pattern variation that does not print is not classified as defective. The usual procedure calls for a very high resolution inspection to find hard defects, looking for their reparation when possible, for a qualification of the repair and then for a series of low resolution soft defects inspection and cleaning cycles until the mask meets specifications. At this point the pellicle frame and the pellicle are attached and another soft inspection is performed. If everything is still in specification the mask is shipped; otherwise the pellicle and its frame are removed and the cycle is performed again.

The inspection has today become, together with the writing, a critical part of the fabrication. The complexity and cost of the inspection tools have actually overtaken that of the writing ones to make it the most expensive equipment in the line. To explain this, it has to be understood that after one or two litho/etch cycles on a mask there must be essentially zero defect larger than the smallest designed feature, otherwise any defect would be replicated multiple times on the wafer reducing the overall chip yield. On a critical mask for the 90 nm node, there must then be zero particle or pattern error larger than 90 nm on the 120 mm × 120 mm active area. This implies inspecting with an

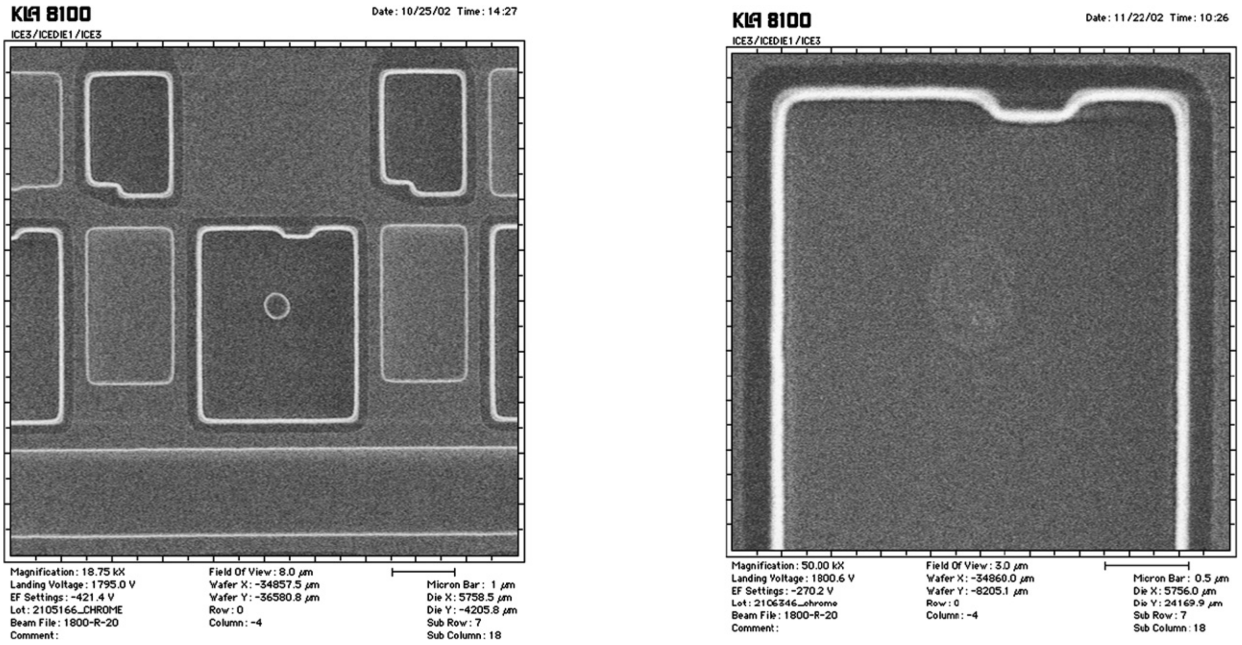

Fig. 10. The left picture shows a clear quartz defect as seen in a SEM. On the right picture the same location is observed after the removal of the defect by an AFM stylus. (Courtesy of Photronics Inc.)

imaging system whose pixel size is of the order of the minimum defect size: in the case of the same 90 nm node it means analysing something like 2×10^{12} pixels. This requirement, moreover, translates in an added defect count per process step considerably lower than that accepted in wafer processing.

These considerations show that the mask should be compared to a wafer-scale chip so the resulting yield is relatively low (rarely above 80% for an established technology and as low as 10% in early development phase). Also one has to note that the yield concept is different from the wafer yield: the completed mask is either acceptable or not, so a yield of 80% means that on average four masks out of five are in specification and the fifth is scrapped and has to be manufactured again with the associated cost and delay. There is a clear interest of having a good understanding of the variations that actually cause problems during printing in order to avoid rejecting ‘good’ masks (affecting mask cost and cycle time) or accepting ‘bad’ masks (affecting chip yield). The study of defect printability is one of the most promising areas to contain mask costs and reduce turnaround time.

The inspection tools are essentially optical tools which analyse the information of both transmitted and reflected lights at different focus positions. The source, for the advanced ones is a DUV laser at ca. 248 nm which is separated in multiple beams which scan the surface. Future tools may need to use an ArF laser to improve the intrinsic resolution. Two modes of operation are possible, one being called ‘die to die’ and the other ‘die to database’. The first one can be applied to masks which have at least two occurrences of an identical pattern (or ‘die’): the information of the scan of one die is compared to the one of the scan of the identical one. Assuming that the probability of having the two different dies with a defect at the same location is negligible, when a difference is observed, the location is logged for further optical reviewing by an operator. The review is performed after the inspection is finished, each defect is classified and the proper repair procedure, if possible and/or necessary, decided.

The ‘die to database’ mode, instead, relies on comparing the information obtained from the image of the die with a database generated at the data preparation. Each difference is again logged and later analysed by an operator. The database is essentially the fractured data file used by the mask writer suitably adapted for the inspection tool and has about the same (large) size. The required data processing capability is very high as the image contrast is very low due to both the nature of the defect (e.g. a fused silica defect in an AAPSM mask as in Fig. 10) and to the resolution limit given by the inspection wavelength. It is necessary to generate an image as close as possible to that logged by the tool using many tool specific parameters. Also adding to the complexity is the fact that the required capture rate is in the order of 99.9% and the false positives need to be kept to a minimum to reduce the operator reviewing time. Again, the use of RET has greatly complicated the inspection task as many of its features are very difficult to distinguish from defects as they have about the same size.

Very rarely the mask is defect free, so the defects are, whenever possible repaired by using specific tools. Extrusion and pindots are usually removed by laser ablation, Focused Ion Beam (FIB) milling, or AFM mechanical nanomachining. In Fig. 10, a fused silica defect in an AAPSM has been removed by an AFM stylus. Intrusion or pinholes are usually repaired by FIB deposition. Extreme care has to be taken in order to restore the correct dimension, line edge continuity and phase and transmission characteristics and advanced RET is again a big problem. FIB is mostly used thanks to its precision and capability of both ablating and depositing material; however, the ion species scattered in the substrate can give rise to phase and transmission changes. For this reason, in the case of AAPSM, the AFM is more used regardless of its lower throughput and the cost of the tips. Electron beams and femtosecond lasers have also been tested but for the moment are at a laboratory stage. YAG lasers have been used in the past for ablation as well as for deposition in association with metallorganic gas atmospheres but their resolution is today insufficient.

In order to insure that the repair is effective, the mask is tested in a special apparatus that determines the aerial image of the mask at a given exposure wavelength and condition. This tool, called AIMS (Aerial Image Measuring System), is essentially a stepper emulator and is routinely used to qualify repairs together with optical and/or SEM observation. Recently this concept has been experimentally applied to defect inspection too. Ultimately it is the final customer who has to provide the test conditions, to decide if the repair is acceptable and to release the acceptance waiver based on the AIMS results.

The fabrication process finishes, after the soft defects inspection and cleaning, with the application of the pellicle frame and of the pellicle and after a final soft defects inspection through the pellicle. The frame is usually made with anodised aluminium with small holes to maintain the pressure balance between the inside and the external ambient. The pellicle is a very thin membrane of organic material which is transparent at the wavelength of use of the reticle and whose main purpose is to keep out of focus any particle that may deposit on the mask. One should note that the use of 193 nm light has generated various problems related to pellicle and frame as the space between pellicle and mask is a perfect chemical reaction chamber and residuals from the pellicle, from the cleaning step or from the glue of the frame can react to produce compounds that crystallise on the mask surface giving rise to a transmission variation. Such effect is usually called ‘haze’ effect and increasing resources are dedicated to change the chemistry of the indicated elements in order to reduce it.

7. The EUV masks

While not a commercial product yet, EUV mask may well be the first real change in the mask industry. Although it is not the only next generation lithography technique requiring very large changes in the mask, we will not discuss here the main contender, namely nano-imprint template, whose application for the critical layers looks even further away from industrial acceptance. EUV masks generate some specific challenges for fabrication and the solution to these issues may impact the industrial introduction of this new technology.

The EUV mask, while maintaining the same sizes, is not transmissive but reflective and the substrate needs to be a very low thermal expansion coefficient material. This is due to the large quantity of energy absorbed by the mask and the requirement that the pattern that is deposited on top of the substrate shows nearly no deformation. Given the wavelength (13.5 nm), the mirror is in fact a Bragg mirror made up of at least 40 bi-layers of Mo and Si whose maximum reflectivity is less than 70% at the 5 to 7 degrees of incidence at which it is used. Although a capping layer is deposed on top of the mirror and below the absorbing material to protect the multilayer stack during process, etching and cleaning processes have to be controlled as not to damage its optical properties. Also the 3D issues will be of larger importance than in current masks. If, at the time of introduction, it will be necessary for imaging to use also phase shifting EUV mask, it will again be necessary to etch into the stack without affecting the reflectivity. Doing so will certainly add to the process criticality.

However, the critical issue for EUV masks is the defectivity: any particle on the surface below the mirror larger than a couple of nm induces a local thickness variation that can totally destroy the imaging. While the top surface defectivity would follow the same requirement of maximum size comparable with feature size, it has to be noted that this implies inspecting and eliminating defects below 32 nm. This is a daunting challenge and no industrial tool exists yet to inspect such defect sizes either for transmissive or reflective masks. The impossibility of using pellicles, as no transparent material exist at EUV wavelength, makes necessary the use of inspection at point of use during the lifetime of the mask. Most of the industrial future of EUV lithography depends on the capability of producing and testing blanks and reticles with an acceptable defect level.

8. Mask industry prospects

The mask manufacturers can be classified in three categories: 'in-house' or 'captive' mask shops of Integrated Device Manufacturers (IDMs), 'in-house' mask shops of foundries and 'merchant' mask shops. They operate with different missions and have different constraint. IDMs' mask shops are usually tasked providing leading edge processes for the internal needs. They are part of the technology capital of the IDM, allow it the earliest possible access to advanced technologies and the possibility of co-developing wafer and mask lithography to reach the best possible compromise. On the other hand, the process developed is usually so specific to the in-house needs that it is very difficult to apply it to fabricate masks for another lithographic process. Foundries mask shops operate much in a similar way but with the added advantage of sharing the cost of a mask set possibly over multiple customers (especially for prototyping and small runs) and easily adapting to the capacity needs. Also they add flexibility in the wafer line as there is a total control of the scheduling of both the wafer manufacturing and the masks. Merchant mask industry, on the other hand, is one of the weakest links of the semiconductor industry. In fact they have to compete among themselves for a reducing pool of customers and have to develop multiple processes according to multiple customer specifications without any guarantee about the actual volume of orders. Although consolidation has taken place and only three global players (2 Japanese ones and one from USA) are left along with some much smaller regional players, especially in Asia, the overall picture is a very fragile one. The overall market in 2005 has been of the order of 3 billion US$ of which about 45% in Japan, and the growth rate forecasts are at less than 5%, well below the total semiconductor market. The merchant maskshops are asked to maintain two very different businesses, one of large volume/low price commodity-like mature technology masks and another of low volume/high price leading edge masks. The first is a low margin business requiring little or no investment, while the second provides a higher margin but high investments with a return on investment (RoI) that barely allows an average 8% expenditure in R&D. The fragility is generated by the fact that every new technology node requires high investment costs while the overall demand is reduced at each node. The reducing demand is the consequence of the increasing cost of new designs and the reduction of the number of customers moving to new nodes. It should be stressed again that the mask business does not track the volume production of chips but only the introduction of new designs or re-designs when there is a change in the layout.

As already pointed out, the period of an 'easy' and cheap access to masks has generated the idea that masks should be a small part of the wafer cost (historically 2 to 4% of overall chip production cost); there is an enormous pressure on price effectively reducing the capability of the mask industry to keep investing. Moreover, the reduction in the volume of high end masks generate a reduction in the number of tools required and consequently an increase of their costs in order to recover the development investments. The net result is an overall reduction of tool suppliers and a shortage in R&D expenditure which generates a delay in the introduction of the much needed tools with improved capability. As an example, today there is only one supplier of inspection tools (outside Japan), one supplier of laser pattern generators, two suppliers of dry etch equipments for mask, one supplier of registration equipment and three suppliers of e-beam mask writers. The small number and size of tool suppliers together with the strong competition may seriously hamper either further mask developments or the timely availability of masks with the required quality. Further consolidation of the mask industry which may contribute to a price increase of the masks and an improvement of the margin of the maskshops, will be of limited benefit to tool suppliers as the tool volume would stay low.

The situation may be partially relieved by any event that can improve the RoI of each part of the supply chain. One possibility is the introduction of direct write tools for the most critical layers. This will slow down the need of further investments by the mask shop while still generating a large business for the less critical layers of the designs. Another way would be through EDA solutions lowering the barrier to generate new designs with a consequent increase of the volume of masks. Those who are more likely to suffer from a possible crisis of the mask industry will be the IDM in the sectors like logic and ASICs where the number of chips per mask set is lower but where a timely availability of the most advanced lithography solutions is critical. They may be the first ones to face the situation that a required lithographic solution is either not available or too expensive. Foundries, especially those with an internal mask shop, and IDMs producing memories and microprocessors will still be able either to mix products on the same reticle or to have volumes large enough to maintain an acceptable reticle cost per chip.

9. Conclusions

The mask industry has been capable of entering the age of sub-wavelength lithography and support the microelectronic industry in its course towards increasing miniaturisation. However its overall shape is weak both technologically and economically, paying some years of low RoI and an increasing gap between R&D needs and actual expenditure. Although each of the multiple technical challenges in itself is not insuperable, the ensemble may be difficult to solve in time and at the right price. The next few years will show if and how the mask industry will still be capable of staying on the roadmap.

Acknowledgements

The author wishes to thank Pierre Sixt and John W. Duff, former colleagues at Photronics, for all the joint work that generated the training material at the base of this paper and also Photronics for permission to using it.

References

[1] B.G. Eynon, B. Wu, Photomask Fabrication Technology, McGraw–Hill, New York, 2005.

[2] SEMI standards SEMI-P1 to SEMI-P6 specify various mechanical and dimensional characteristics of masks. They are available via SEMI.

[3] B. Kasprowicz, S. Bryan, R. Priestley, Evaluation of new mask materials for improved lithography performance, in: Optical Microlithography XIV, Proc. SPIE 4346 (2001) 4346–4383.

[4] SEMI P39-1105 OASIS—Open Artwork System Interchange Standard.

[5] E. Robert, et al., Strategies of optical proximity correction dedicated to chromeless phase lithography for 65 and 45 nm node, in: Optical Microlithography XVIII, Proc. SPIE 5754 (2005) 476–487.

[6] C.J. Progler, A. Borna, D. Blaauw, P. Sixt, Impact of lithography variability on statistical timing behaviour, in: Design and Process Integration for Microelectronic Manufacturing II, Proc. SPIE 5379 (2004) 101–110.